Semi-Supervised Learning for Named Entity Recognition Using Weakly Labeled Training Data

Atefeh Zafarian
HLT Lab.
Dept. of Computer Eng. and IT
Amirkabir University of Technology, Tehran, Iran
zafarian@aut.ac.ir

Ali Rokni
HLT Lab.
Dept. of Computer Eng. and IT
Amirkabir University of Technology, Tehran, Iran
alirokni@eecs.wsu.edu

Shahram Khadivi
HLT Lab.
Faculty of Computer Eng. and IT
Amirkabir University of Technology, Tehran, Iran
khadivi@aut.ac.ir

Sonia Ghiasifard
HLT Lab.
Dept. of Computer Eng. and IT
Amirkabir University of Technology, Tehran, Iran
ghiasifard@aut.ac.ir

***Abstract*— The shortage of the annotated training data is still an important challenge to building many Natural Language Process (NLP) tasks such as Named Entity Recognition. NER requires a large amount of training data with a high degree of human supervision whereas there is not enough labeled data for every language. In this paper, we use an unlabeled bilingual corpora to extract useful features from transferring information from resource-rich language toward resource-poor language and by using these features and a small training data, make a NER supervised model. Then we utilize a graph-based semi-supervised learning method that trains a CRF-based supervised classifier using that labeled data and uses high-confidence predictions on the unlabeled data to expand the training set and improve efficiency of NER model with the new training set.**

***Index Terms*— Named entity Recognition, Bilingual parallel corpora, graph-based semi-supervised learning**

I. INTRODUCTION

Named entities are phrases that contain the names of person, organization and location and Named entity recognition is an important task of natural language process to identifies occurrences of the phrases as belonging to particular categories of Named entities. Many NLP tasks require a large amount of training data with a high degree of human supervision. Unfortunately, all languages have not enough labeled data for many languages building. Although there are powerful supervised approaches [19], [13], [4], accessing to a sufficient amount annotated data is not only a key success factor of these systems in resource-rich languages but also a challenging factor in resources poor languages. This fact motivates semi-supervised and unsupervised approaches for NER in resource-poor languages. In this paper, we want to transfer the supervised information from resource-rich language toward resource poor languages in bilingual parallel corpora. This work provides useful features for labeling data on resource-poor language.

To have an efficient system in NLP task, we extract useful features from bilingual corpus and make a NER supervised model, then we utilize a semi-supervised learning method on resource-poor language to expand the training set and improve efficiency of NER model. For this goal, at first, we make a powerful system for resource-rich language, and then transfer named entities from resource-rich language toward resource-poor language by using the alignment model and transliteration named entity. We extract useful features from this transferring and make a NER supervised model. In continuous, we manually annotate named entities on a small part of the bilingual corpora and build a small training set and train the NER supervised classifier on the training set. Finally, we use a graph-based semi-supervised learning that leverages both labeled and unlabeled data to improve our NER model performance. This method iteratively trains a CRF-based supervised classifier using that labeled data and uses high-confidence predictions on the unlabeled data to expand the training set. In this work, at first, we train a CRF-based supervised classifier on labeled data and use it to decode unlabeled data, then create a graph where the nodes are both labeled and unlabeled samples and extract a set of NER features for each sample and connect similar nodes based on these features. The state posteriors on unlabeled data are then smoothed using the graph regularizer and knowledge graph is combined with the knowledge supervised method and best label for any unlabeled sample is generated. Finally, we extract high-confidence sentence on unlabeled data to expand the training set and train NER model with the new training set.

The proposed method is language-independent and we can use it for each language. In this paper, the experiments are conducted in Persian language which suffers from the shortage of NLP resources. There are two corpora for Persian language, bilingual corpora and IUST corpus. The bilingual corpora contains 656938 English-Persian sentence pairs that is provided by [9] and a part of this corpus is labeled manually for generating NER test set. IUST corpus is a Persian NER corpus with more than 180000 words that is labeled manually for the NER tasks and is divided into a training (90%) and test (10%) sets.

When we train a NER supervised classifier on IUST training corpus and test on IUST test corpus, F1 score is more than 80%, whereas when we test that on the bilingual test set, F1 score is lower than 41%. According to the method described, we obtain a 26% increase in F1 score on the bilingual test set.

978-1-4799-8818-1/15/$31.00 ©2015 IEEE

II. SUPERVISED LEARNING

In this section, we make a supervised learner model on IUST training data. Whereas there are a lot of methods for making the learner model in NLP systems, CRF method achieve a high performance in the Persian NER system so we use CRF-based Stanford named entity tagger to tag named entity. CRF-based learner models in NER tasks use common features such as the order of the CRF, the assigned class of the word, the word itself, its n-gram and the prior and posterior word. We investigated the different features of a Persian NER system and identified the best set of features for a Persian NER system, then we train the Stanford NER tagger on the IUST training set and test it on IUST and bilingual test set. The results of these experiments are shown in table 1,2. These results show when training and test set are selected from the IUST corpus, we obtain an efficient result, whereas, when test set is selected from bilingual corpus, the quality of model is extremely reduced.

Here, we describe the main causes of the poor result on the bilingual test set. The supervised methods usually need a large number of training data, whereas the size of the IUST training corpus is small also when the domain of training and test are different, supervised methods don't perform well. In the second experiment, the domain of training and test sets are different and also test set come from a corpus with a diverse range of topics. There are also human errors in both of corpus like incorrect tagging in the training set, spelling errors and different translation of one word in the bilingual test set.

Table 1. Results IUST model on its test set

Entity	Precision	Recall	F1-score
LOCATION	95.10	73.65	83.01
ORGANIZATION	92.28	77.95	84.51
PERSON	91.50	64.52	75.68
TOTAL	93.17	71.60	80.97

Table 2 . Results IUST model on bilingual test set

Entity	Precision	Recall	F1-score
LOCATION	94.87	41.11	57.36
ORGANIZATION	33.33	4.55	8.00
PERSON	54.55	12.00	19.67
TOTAL	83.02	27.16	40.93

III. PROPOSED METHOD

Our method includes 6 steps:

- Transferring the known named entities from the English side of the bilingual corpus toward Persian text by using Stanford tagger and an alignment output of GIZA++.
- Limiting the number of probable translations by measuring the phonetic distance of the words.
- Generating the tagger model by using the alignment probability, phonetic distance, the transferred tag, available model tag and its percent confidence.
- Apply graph-based semi supervised learning for expanding training set.
- Learning the tagger model by the new training set.
- Evaluation of the model.

A. *Transferring Named Entities*

At the first stage, we tag the English sentences of the bilingual parallel corpus by using the Stanford named entity tagger. The possible tags for the named entity are PERSON, LOCATION, ORGANIZATION. If the word has not any of these 3 states, it will be tagged as an O. Then we apply GIZA++ on this parallel corpus. GIZA++ is the part of the Moses tool that is utilized for the machine translation. By using GIZA++ output, we can gain the equivalent probability of each word in the English Sentence to one word in the Persian sentence and vice versa.

B. *Compute of Phonetic Distance*

The most named entities like people's name, location or organizations are rooted in the source language and have not the semantic translation in English language, typically these words are translated like a Persian word as the way they are pronounced. For this purpose, the phonetic distance Of the two words is a suitable criteria to identify the applied alignment accuracy by GIZA++. In this section, we will introduce a new distance for the two words and name it, the weightful phonetic distance. This distance is different from other distances like levenshtein. The first difference as the name shows, this distance is not about the equality of the letters, but it's about the equivalence of the pronunciation process of two words. Secondly, unlike the levenshtein distance, this distance is one to many. For example, two "sh" letters in the English words are equivalent to "ش " in Persian word. The last difference is weighting of the mappings. Mapping types are described in table 3:

Table 3. Types of mappings and these costs

Mapping type	Cost
English non-vowel to Persian non-vowel pairs	0
English non-vowel to Persian non-vowel not pairs	2
English vowel to Persian vowel	0
English vowel to Persian non-vowel	1
Removing non-vowel English/ Persian letter	2
Removing vowel English/ Persian letter	1

As in the Persian language the short vowels are not written, in the English equivalent of the Persian word, maybe these vowels exist. So the omission of these vowels is done with lower cost. On the other hand the omission of the consonant letters is done with higher cost. For measuring the distance at each point, we will use the recursion definition according to Eq. 1. In this definition, e and f are English and Persian words respectively.

$$\begin{aligned} &d(i, j) = \min(d(i-1), j) + \text{remove}(e_i), \\ &d(i, j-1) + \text{remove}(f_j), \\ &d(i-1, j-1) + \text{replace}(e_i, f_j), \\ &d(i-2, j-1) + \text{replace}(e_{i-1}e_i, f_j). \end{aligned} \quad (1)$$

C. Make Learner Model

In this section, we label manually a small part of the bilingual corpus, including 10716 words in 159 sentences. As the manual corpus is small, for keeping the comprehensiveness of the corpus, we select data in a way that each of the possible cases has at least 100 samples at the training corpus. Then, we make a CRF-based NER model and train it on labeled data that we label manually.

Normally, the learner models based on conditional random field use of common features. The common features include the order of conditional random field, the assigned class to the word, the word itself, its n-gram and the prior and posterior word. We will also add other features to the model for learning. On one hand, the first model that had not suitable efficiency on the bilingual corpus will be applied on the transferred data. The output of this application is the assigned tags and their percent confidence.

D. Feature Selection

In addition, extracted features from the sentence, the information gained from transferring process, will provide us with the definition of new features. The new features are as followed:

- The transferred tag from the English corpus
- The alignment probability of the tagged Persian word to the corresponding English word.
- The normalized phonetic distance of two English and Persian words.
- Tagging the first model on the word.
- Confidence of the first model tag.

Although, by defining a simple heuristic function, we can use somewhat from the effect of these parameters in determining the final tag. We attempted by using a systematic process, give the combination process of the parameters to a learner model. The learning process will be done by the Stanford named entity tagger. At this stage, learning will be done by the added features on the tagged corpus. The result of running this model on bilingual test set is shown in table 4. In this section, we obtain a 23% increase in F1 score on the bilingual test set.

Table4 . Results learner model on bilingual test set

Entity	Precision	Recall	F1-score
LOCATION	77.92	74.07	75.95
ORGANIZATION	64.71	31.43	42.31
PERSON	71.05	50.00	58.70
TOTAL	74.24	57.65	64.9

E. GRAPH-BSED SEMI-SUPERVSED LEARNING

In the previous sections, by using the available features in bilingual corpus and one small labeled corpus, we created a learner model for NER and obtained an acceptable result on the bilingual test set. Whereas, there is a large number of unlabeled data that we can use it for improving NER model. For this goal, we apply a graph based semi-supervised learning and use it to tag unlabeled data. Then, we extract sentences with the reliable tags and add those to the training set for expanding the training set.

F. Graph Construction

Graph construction is an important step in graph-based SSL. There are three main steps in the graph construction procedure, determine vertices, determine the similarity function and select edge between vertices. Vertices in graph consist of all the word 3-grams (types) that have occurrences at least once in the labeled sentences or only in the unlabeled sentences. We keep information about the beginning and end of sentences by adding phrases “start” in begin and “end” in the end of sentences. Each vertex in the graph is unique. Types u and v are linked by an edge of Weight W_{uv} that W_{uv} is determined by a similarity function. We define the similarity function based on existing features in named entities. The NER is mostly determined by its local context so we extract a set of n-gram features according to table 5. Also named entities are often preceded by a set of prefixing title words that we use these words as features in similarity function. Some of these words are shown in table 5.

Table5. Features table

Description	Feature
Trigram	$x2\,x3\,x4$
Left Word	$x2$
Right Word	$x4$
Left Word	$x1\,x2$
Right Word	$x4\,x5$
Prefix of position name	City, Country, …
Prefix of organization name	Organization, university, ..
Prefix of person name	Mr, Ms, …

For computing similarity between two nodes, we compute the vector of Point-wise Mutual Information (PMI) values between each type and each of the features and then use the cosine distance between those PMI vectors as our similarity function.

G. Semi supervised learning algorithm

At first, we train a CRF-based supervised tagger on labeled data and compute the marginal probabilities over Named Entity (NE) tags for each word item in labeled and unlabeled data. By knowing word items, we know "type" and its CRF probability. Each type of graph is unique, if the type occurs more than once in the data, we calculate the mean of its CRF probability as type probability. We iterate over sentences and their types and retrieve features of each type, then we compute PMIs of each node and connect similar nodes. After this step each node has a set of neighbors and a CRF probability.

We propagate graph with the same method as [16]. In this work, by minimizing following convex objective, the type-level marginals are smoothed. In this equation, we obtain the type marginal for each type as minimizing Eq. 2.

$$C(q) = \sum_{u \in V_l} \|r_u - q_u\|^2 + \mu \sum_{u \in V, v \in N(i)} W_{uv} \|q_u - q_v\|^2 + \vartheta \sum_{u \in V} \|q_u - U\|^2$$
$$s.t. \sum_y q_u(y) = 1 \quad \forall u \ \& \ q_u(y) \geq 0 \quad \forall u, y \qquad (2)$$

Eq. 2, contains 3 terms. Term 1 penalize labeled nodes that received different label with its original label where r_u is the empirical marginal label distribution for node u. Term 2 penalizes neighboring nodes that have different label distribution. Term 3 regularizes the label distribution toward the uniform U over all possible labels y.

In the Viterbi decode, we interpolate the type marginals computed in the previous step with the original CRF marginals by a linear combination and compute a new marginal for each word item in unlabeled sentences. After this step, each word item in unlabeled data has a label that we can use it for retraining CRF. We re-estimate CRF parameter with the new train set and retrain CRF with new parameter and iterate algorithm until convergence.

The described method propagates some errors in the output. We investigate errors and propose some solutions to solve these problems. Types contain "start" and "end" connect with one another with high weight, while not necessarily similar, we don't use "start" or "end" when compute PMI. Center words in some types are prepositions, signs or verbs, an incorrect labeling in these types propagates error while a NER graph doesn't need these types. So we remove these from the node list.

Also, in Viterbi decode phase, we use a word threshold parameter and remove each sentence that has at least any word which the confidence of major NE tag is lower than this

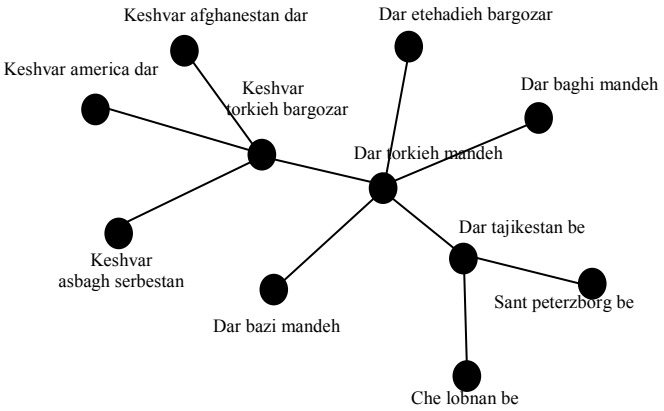

Figure 1. graph with center word 'torkieh' and their local neighborhoods

threshold, also we compute a sentence confidence score using following Eq. 3.

We use $\max(10, sentencelength)$ because some of the sentences were unexpectedly short and contains a long NE and these abnormal sentences score a lot. We propagate unlabeled sentences that their score is more than sentence threshold. By doing this job, we only use sentences with high confident words and prevent error propagation of low confident words.

$$sentence\ confidence = \frac{\sum_{word \in i} word\ confidence}{\max(10, sentencelength)} \qquad (3)$$

Also, we apply two ideas to increase precision and recall in our model. According to the results reported in table 1, CRF-based NER model obtained an acceptable result in precision measure. Then, we increase precision by prioritizing node's CRF tag over newly decoded tag when its CRF tag is NE and its new tag is O. Also, we assign parts of speech to each word in unlabeled data by an existing POS tagger [14] for Persian language and increases recall by using the POS tag. As a rule of thumb, NEs are from nouns and therefore, we remove the NE tag from words that have not noun root POS (such as verb).

H. An Example of a Graph Operation

CRF-based supervised tagger assign a label to each word based on local features whereas local features alone can't determine the proper label for each word in a NER system also the efficiency of these methods is dependent on the source domain and they are poor at recognizing when words occur in a limited time in the source domain. Self-training methods just reinforce the knowledge of supervised method, whereas the graph-based methods use knowledge gained from the similarity of nodes to recognize labels.

In this section, we show the advantage by a sample. The word "torkieh" don't occur in the source domain, whereas it occurs in unlabeled data in trigrams "*dar torkieh mandeh*" and "*keshvar torkieh bargozar*". The correct label for "*torkieh*" is LOC, a CRF-based self-training method labels it as LOC for "*keshvar torkieh bargozar*" and a non-proper noun for "*dar torkieh mandeh*". Figure 1 shows a part of the graph. The figure is the neighborhoods of a subset of the vertices with the center word "*torkieh*", node "*dar torkieh mandeh*" connect to "*dar bazi mandeh*","*dar baghi mandeh*" and "*keshvar torkieh bargozar*" with the high weight. "*dar bazi mandeh*" and "*dar baghi mandeh*" are non-proper noun in labeled data. Because of local features in "*dar torkieh mandeh*" are close to "*dar bazi mandeh*" and "*dar baghi mandeh*", self-training method label "*dar torkieh mandeh*" as a non-proper noun with high probability, whereas connecting "*dar torkieh mandeh*" and "*keshvar torkieh bargozar*" in the graph, propagates a correct current in the graph and increases the probability of LOC for "*dar torkieh mandeh*".

IV. Experiments and Results

At first, we trained a CRF-based supervised model (model 1) on IUST training set and tested it on bilingual test set and obtained an F1 score of 41% for this experiment. The results of these experiments in table 6 are shown when the domain of training and test are different, supervised methods don't perform well.

Table 6. Results of model (1) on bilingual test set

Entity	Precision	Recall	F1-score
LOCATION	94.87	41.11	57.36
ORGANIZATION	33.33	4.55	8.00
PERSON	54.55	12.00	19.67
TOTAL	83.02	27.16	40.93

So, we created a learner model (model 2) in section III by using the features gained from transferring named entities from English to Persian language and the small training data, and obtained an acceptable result on the bilingual test set, in this step, we obtained a 23% increase in F1 score on the bilingual test set. The results of these experiments are shown in tables 7.

Table 7. Results model (2) on bilingual test set

Entity	Precision	Recall	F1-score
LOCATION	77.92	74.07	75.95
ORGANIZATION	64.71	31.43	42.31
PERSON	71.05	50.00	58.70
TOTAL	74.24	57.65	64.9

As a result by using a bilingual corpus, we can increase drastically the efficiency of identification of the named entities in the bilingual test set. Another visible point at this experience is the suitable efficiency method for recognizing and identifying the location and person names, because the person and location names are usually transliterated in bilingual corpora. So the phonetic distance is a good criteria for mapping word to its translation. For the names of companies and organizations, the conditions are a little bit different. For example "sazman melall" is the Persian equivalent of the phrase "united nations" whereas their phonetics aren't same. This case leads to making mistake in transliteration phase. But in spite of this problem the efficiency of this model in recognizing organization's names is again better than that of first model.

Finally, we used a graph-based semi-supervised learning method and generated a new train set. This new set contains labeled data and extracted sentences with the reliable label from unlabeled data. We trained the NER model by new training corpus and tested on the bilingual test set. In this section, we also obtained a 3% increase in F1 score than previous model. The results of these experiments are shown in table 8.

Table 8. Results model (2)+semi-supervised learning on bilingual test set

Entity	Precision	Recall	F1-score
LOCATION	75.61	76.54	76.07
ORGANIZATION	61.11	31.43	41.51
PERSON	72.34	62.96	67.33
TOTAL	72.79	62.94	67.51

We also compare our method with a self-training method [17]. In this iterative algorithm, we train CRF-based NER model on the initial training set and compute the marginal probabilities over unlabeled data. We extract high-confidence sentence on unlabeled data and add those to the initial training set and iterate algorithm. As before, we use threshold parameters for words and sentences. We train our NER model on this training set and test on bilingual test set. Table 9 shows the result of this experiment.

Table 9. Results self-training model on bilingual test set

Entity	Precision	Recall	F1-score
LOCATION	73.85	59.26	65.75
ORGANIZATION	41.67	28.57	33.90
PERSON	67.35	61.11	64.04
TOTAL	65.94	53.53	59.09

The results of table 9 show that not only the use of graph-based semi-supervised have been effective in improving outcomes, but also the learner model developed in section III

obtained a better result than the self training method. Using the graph structure and trigram phrases in semi-supervised method and extracted features from bilingual corpora made our NER system different and more effective than self-training.

V. Related Work

There are many different approaches in NER by supervised learning methods, include HMM [2], ME [3], SVM [1] and CRF [11]. The common features of these approaches are that these need to a large and comprehensive set of labeled data. Unfortunately, all languages have not enough labeled data for many languages building, therefore the resource-poor languages use to the bilingual corpus for transferring information from resource-rich language toward resource-poor language in NLP task. There is a large amount work with this method for NER. For example, [18] used an English-France bilingual corpora and [7] used an English-Chinese bilingual corpora for transferring named entity from English to target language. [8] presented an effective integrated approach that can improve the extracted named entity translation dictionary and the entity annotation in a bilingual training corpus. [12] produced named entity annotated corpora in foreign languages by transferring knowledge from English Wikipedia. [10] also used a method similar for NER. The lack of sufficient labeled training data is the reason for the recent attention towards semi-supervised learning methods. [5] and [6] presented algorithms to automatically discover NEs from untagged corpora with minimal supervision. [20] used a semi-supervised learning based on a Gaussian random field that a similarity graph was used to propagate information between labeled and unlabeled data. Also [15] used a method graph-based semi-supervised learning to label unlabeled data.

VI. Acknowledgment

The tagged corpus is one of the essential needs in supervised learning methods. Providing this corpus by human is a too much expensive and time consuming work. On one hand the efficiency of supervised method will decrease massively along with changing the domain of texts. So that, we get a poor result When, we train the Stanford NER tagger on the IUST training set and test it on bilingual test set. At this paper, we showed that by using a bilingual corpus and transliteration technique, can make a model carefully with low costs on the bilingual corpora. We also showed that by using a graph-based semi-supervised method, can extract the reliable sentences from an unlabeled data to expand the training set and enhance the quality of our NER model.

References

[1] M.Asahara, Y.Matsumoto, "Japanese named entity extraction with redundant morphological analysis", In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1 (pp. 8-15). Association for Computational Linguistics, May 2003.

[2] D.Bikel, S.Miller, R.Schwartz, and R.Weischedel, "Nymble: a high-performance learning name-finder", In Proceedings of the fifth conference on Applied natural language processing (pp. 194-201). Association for Computational Linguistics, March 1997.

[3] A.Borthwick, J.Sterling, E.Agichtein, and R.Grishman, "Description of the MENE Named Entity System as used in MUC-7", In Proc. Seventh Message Understanding Conference, 1998.

[4] H.Chieu, and H.Ng, "Named entity recognition: a maximum entropy approach using global information", In Proceedings of the 19th international conference on Computational linguistics-Volume 1 (pp. 1-7), Association for Computational Linguistics, August 2002.

[5] S.Cucerzan, and D.Yarowsky, "Language independent named entity recognition combining morphological and contextual evidence", In Proceedings of the 1999 Joint SIGDAT Conference on EMNLP and VLC (pp. 90-99), June 1999.

[6] M.Collins, Y.Singer, P.Avenue, and F.Park, "Unsupervised Models for Named Entity Classification" , 100–110.

[7] R.Fu, B.Qin, AND T.Liu, "Generating Chinese Named Entity Data from a Parallel Corpus", In IJCNLP (pp. 264-272), 2011.

[8] F.Huang, and S.Vogel, "Improved named entity translation and bilingual named entity extraction", In Multimodal Interfaces, 2002. Proceedings. Fourth IEEE International Conference on (pp. 253-258), IEEE, June 1999.

[9] F.Jabbari, S. Bakhshaei SM.Mohammadzadeh Ziabary and S. Khadivi, "Developing an Open-domain English-Farsi Translation System Using AFEC: Amirkabir Bilingual Farsi-English Corpus", Association for MachineTranslation in the Americas (AMTA 2012) (2012).

[10] J.Nothman, JR.Curran, and T.Murphy, "Transforming Wikipedia into named entity training data", In Proceedings of the Australian Language Technology Workshop (pp. 124-132), December 2008.

[11] A.McCallum, and W.Li, "Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons", In Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4 (pp. 188-191). Association for Computational Linguistics, May 2003.

[12] AE.Richman, P.Schone, and FGG.Meade, "Mining Wiki Resources for Multilingual Named Entity Recognition" ,1–9, June 2008.

[13] B.Settles, "Biomedical named entity recognition using conditional random fields and rich feature sets", In Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (pp. 104-107), Association for Computational Linguistics, August 2004.

[14] Z.Shakeri, N.Riahi, and S.Khadivi, "Preparing an accurate Persian POS tagger suitable for MT".

[15] A.Subramanya, and JA.Bilmes, "Entropic graph regularization in non-parametric semi-supervised classification", In Advances in Neural Information Processing Systems (pp. 1803-1811), 2009.

[16] A.Subramanya, S.Petrov, and F.Pereira, "Efficient graph-based semi-supervised learning of structured tagging models", In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (pp. 167-176), Association for Computational Linguistics, 2010, October 2010.

[17] D.Yarowsky, "Unsupervised word sense disambiguation rivaling supervised methods", In Proceedings of the 33rd annual meeting on Association for Computational Linguistics (pp. 189-196). Association for Computational Linguistics, June 1995.

[18] D.Yarowsky, G.Ngai, and R.Wicentowski, "Inducing multilingual text analysis tools via robust projection across aligned corpora", In Proceedings of the first international conference on Human language technology research (pp. 1-8), Association for Computational Linguistics, March 2001.
[19] G.Zhou, and J.Su, "Named entity recognition using an HMM-based chunk tagger", In proceedings of the 40th Annual Meeting on Association for Computational Linguistics (pp. 473-480), Association for Computational Linguistics, July 2002.
[20] X.Zhu, Z.Ghahramani, and J.Lafferty, "Semi-supervised learning using gaussian fields and harmonic functions", In ICML (Vol. 3, pp. 912-919), August 2003.